

Natural Language Processing

Project 2

Alizée Pace
alizée.pace@ai.ethz.ch
29.03.2021

Motivation: Give you a taste for the usefulness of NLP in medicine!

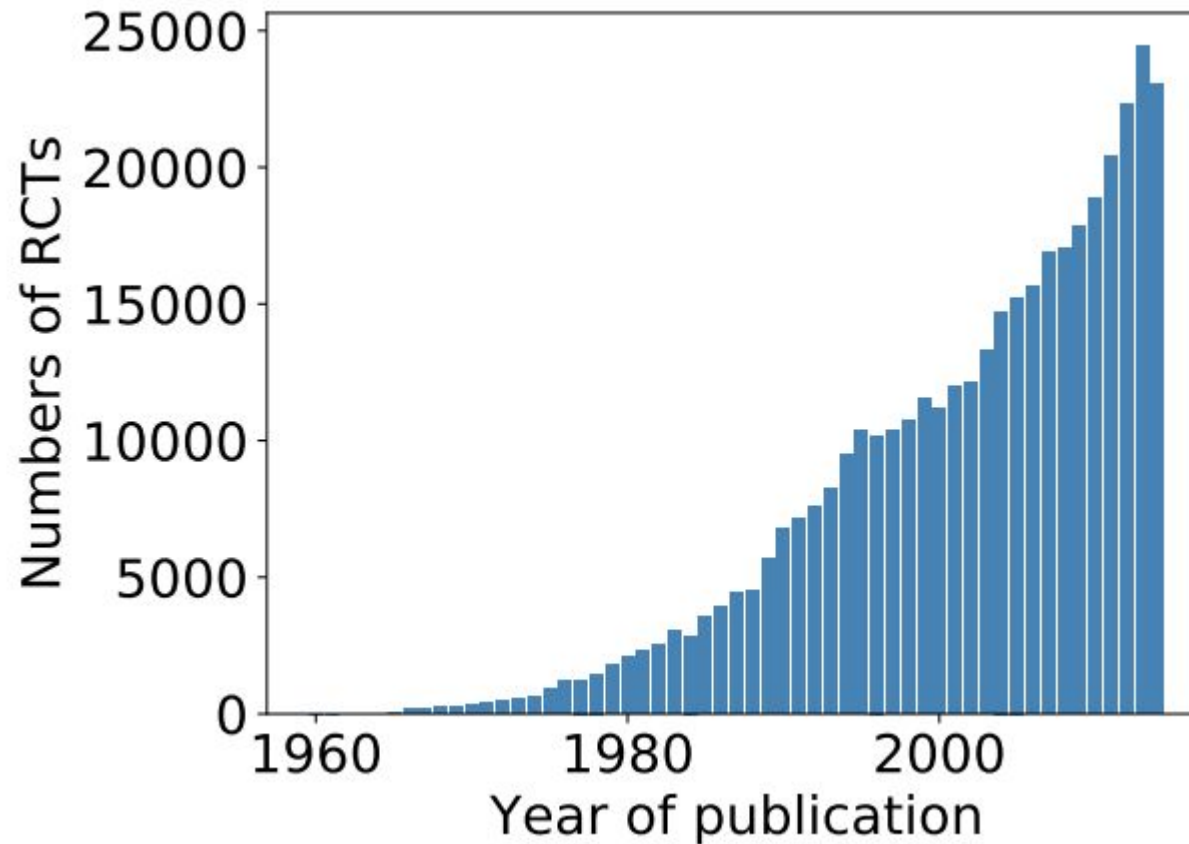


Language: succinct, efficient means of transmitting information.

- Endless sources of information in medical text:
 - 🏥 patient records & letters
 - 👩 doctor notes
 - 💬 conversation transcripts
 - 💰 insurance claims
- Privacy and ethics issues with most of these. Let's focus on publicly available, non-identifiable data:
 - 📚 medical literature
 - 📰 journalism
 - 🧠 social media

Source: Silvia Natalia on vecteezy.

The medical literature is huge...



- Randomized Controlled Trials (RCTs): best source of medical evidence.
- >1M published so far: **Challenging to efficiently parse the existing literature!**

Source: Dernoncourt & Lee, IJCNLP 2017.

Some attempts to facilitate literature parsing

- literature reviews
- graphical abstracts
- clear outlining of results
- structured abstracts (→ ~50% of literature)

Randomized controlled trial hepatocellular carcinoma

Bo-Heng Zhang¹, Bing-Hui Yang, Zhao-You Tang

Affiliations + expand

PMID: 15042359 DOI: [10.1007/s00432-004-0552](https://doi.org/10.1007/s00432-004-0552)

Abstract

Purpose: Screening for hepatocellular carcinoma (HCC) in urban Shanghai, China, where there is no conclusive evidence that screening may be effective. The purpose of this study was to assess the effect of screening on HCC mortality.

Methods: This study included 18,816 people, aged 40-74 years, with a history of chronic hepatitis in urban Shanghai, China. They were randomized to screening (9,373) or control (9,443) group. Control group participants received health-care facilities. Screening group participants received ultrasonography examination every 6 months. Screening group participants had been offered screening up until December 1998. The primary outcome measure was HCC mortality.

Results: The screened group completed 58.2 percent of the screening offered. The screened group was compared to the control group, the number of HCC was 52 (60.5%) versus 0; small HCC 39 (45.3%) versus 5 (7.5%); 1-, 3-, and 5-year survival rate 65.9%, 52.6%, 46.5%, respectively. Thirty-two people died from HCC in the screened group versus 0 in the control group, and the HCC mortality rate was significantly lower in the screened group (being 83.2/100,000 and 131.5/100,000, respectively, with a mortality rate ratio of 0.98).

Conclusions: Our finding indicated that biannual screening reduced HCC mortality.

Pain exposure physical therapy (PEPT) compared to conventional treatment in complex regional pain syndrome type 1: a randomised controlled trial

Karlijn J Barnhoorn¹, Henk van de Meent², Robert T M van Dongen³, Frank P Klomp⁴, Hans Groenewoud⁵, Han Samwel⁶, Maria W G Nijhuis-van der Sanden⁷, Jan Paul M Frölke⁸, J Bart Staal⁹

Affiliations + expand

PMID: 26628523 PMID: [PMC4679993](https://pubmed.ncbi.nlm.nih.gov/26628523/) DOI: [10.1136/bmjopen-2015-008283](https://doi.org/10.1136/bmjopen-2015-008283)

Free PMC article

Abstract

Objective: To compare the effect of pain exposure physical therapy (PEPT) with conventional treatment in patients with complex regional pain syndrome type 1 (CRPS-1).

Setting: The study was conducted in a specialised outpatient clinic.

Participants: 56 adult patients with CRPS-1.

Interventions: Patients received either PEPT or conventional treatment for 12 weeks.

Measurements: Outcome measures included Pain Catastrophising Scale (PCS), Tampa Scale of Kinesiophobia (TSK-11), and EuroQol-5D.

Results: The primary outcome was the change in PCS score. The PEPT group showed a significantly greater reduction in PCS score compared to the conventional treatment group.

Conclusions: PEPT is more effective than conventional treatment in reducing pain catastrophising in patients with CRPS-1.

Survival outcomes following laparoscopic versus open D3 dissection for stage II or III colon cancer (JCOG0404): a phase 3, randomised controlled trial

Seigo Kitano¹, Masafumi Inomata², Junki Mizusawa³, Hiroshi Katayama³, Masahiko Watanabe⁴, Seiichi Yamamoto⁵, Masaaki Ito⁶, Shuji Saito⁷, Shoichi Fujita⁸, Fumio Konishi⁹, Yoshihisa Saida¹⁰, Hirotoshi Hasegawa¹¹, Tomonori Akagi¹, Kenichi Sugihara¹², Takashi Yamaguchi¹³, Tadahiko Masaki¹⁴, Yosuke Fukunaga¹⁵, Kohei Murata¹⁶, Masazumi Okajima¹⁷, Yoshihiro Moriya⁵, Yasuhiro Shimada¹⁸

Affiliations + expand

PMID: 28404155 DOI: [10.1016/S2468-1253\(16\)30207-2](https://doi.org/10.1016/S2468-1253(16)30207-2)





Abstract

Background: Although benefits of laparoscopic surgery compared with open surgery have been suggested, the long-term survival of patients undergoing laparoscopic surgery for colon cancer requiring Japanese D3 dissection remains unclear. We did a randomised controlled trial to compare the non-inferiority of laparoscopic surgery to open surgery.

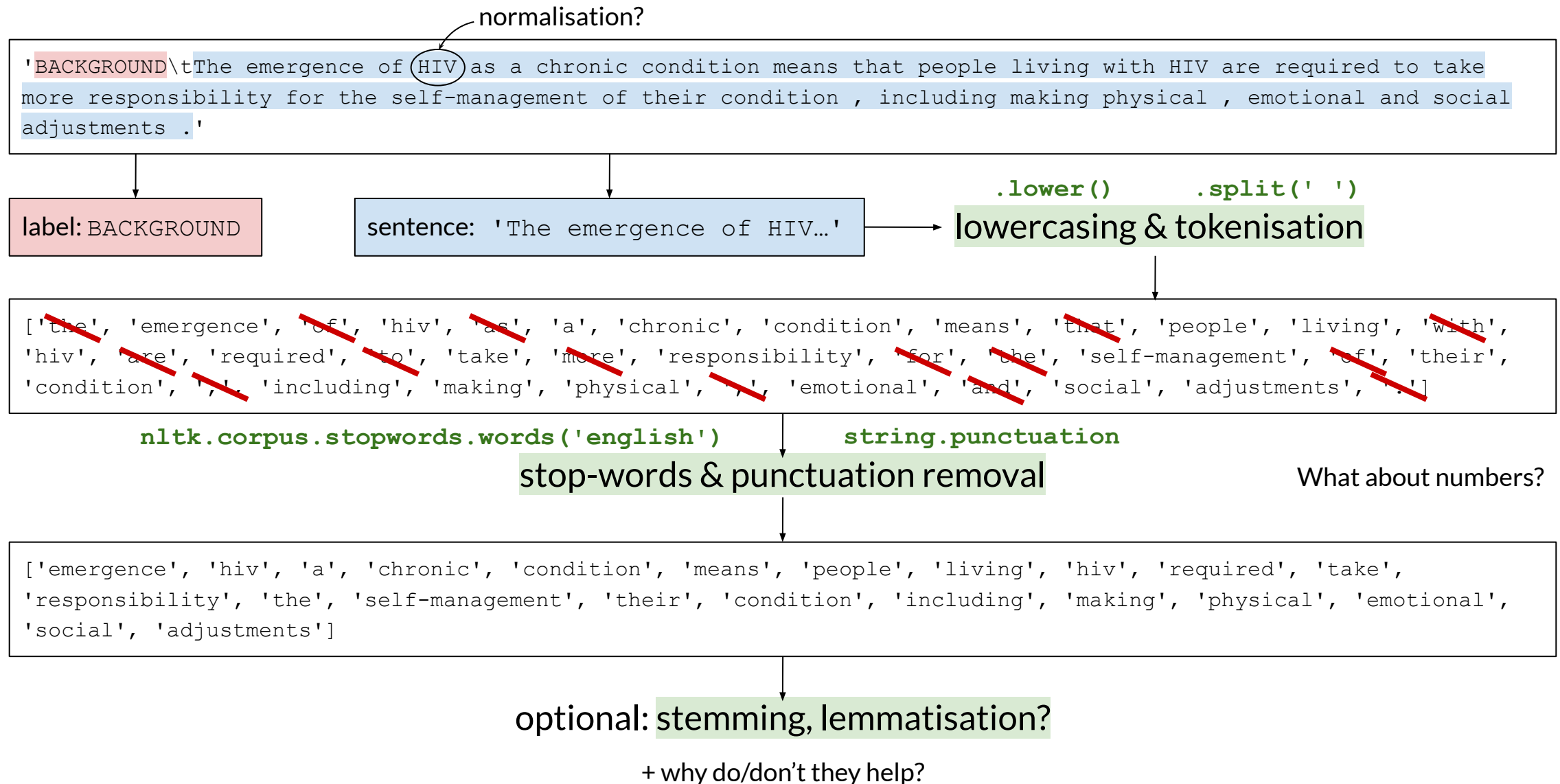
Methods: We did an open-label, multi-institutional, randomised, two-arm phase 3 trial in 3 hospitals in Japan. Patients aged 20-75 years who had histologically proven colon cancer; T3 or deeper lesions without involvement of other organs, node stages N0-2, and metastasis stage M0; and tumour size ≤ 10 cm were included. Only accredited surgeons did surgery as an operator or instructor. Patients were randomly assigned (1:1) preoperatively to undergo D3 resection either by an open or a laparoscopic route, via phone call or fax to the Japan Clinical Oncology Group (JCOG) Data Center. Randomisation used a minimisation method with a biased-coin assignment according to tumour location (caecum, ascending vs sigmoid, rectosigmoid) and institution. The primary endpoint was overall survival and was analysed by intention to treat. The non-inferiority margin was set at 1.366. This study is registered with UMIN Clinical Trials Registry number C000000105, and ClinicalTrials.gov, number [NCT00147134](https://clinicaltrials.gov/ct2/show/study?term=NCT00147134).

Findings: Between Oct 1, 2004, and March 27, 2009, 1057 patients were randomly assigned to either open surgery (n=528) or laparoscopic surgery (n=529). 5-year overall survival was 90.4% (95% CI 87.5-92.6) for open surgery and 91.8% (89.1-93.8) for laparoscopic surgery. Laparoscopic D3 surgery was not non-inferior to open surgery for overall survival (HR 1.06, 95% CI 0.79-1.41; P_{non-inferiority}=0.073). 65 (13%) patients in the open surgery group and 53 (10%) patients in the laparoscopic surgery group had grade 2-4 adverse events. Grade 2-4 adverse events included diarrhoea (15 [3%] in the open surgery group vs 14 [3%] in the laparoscopic surgery group).

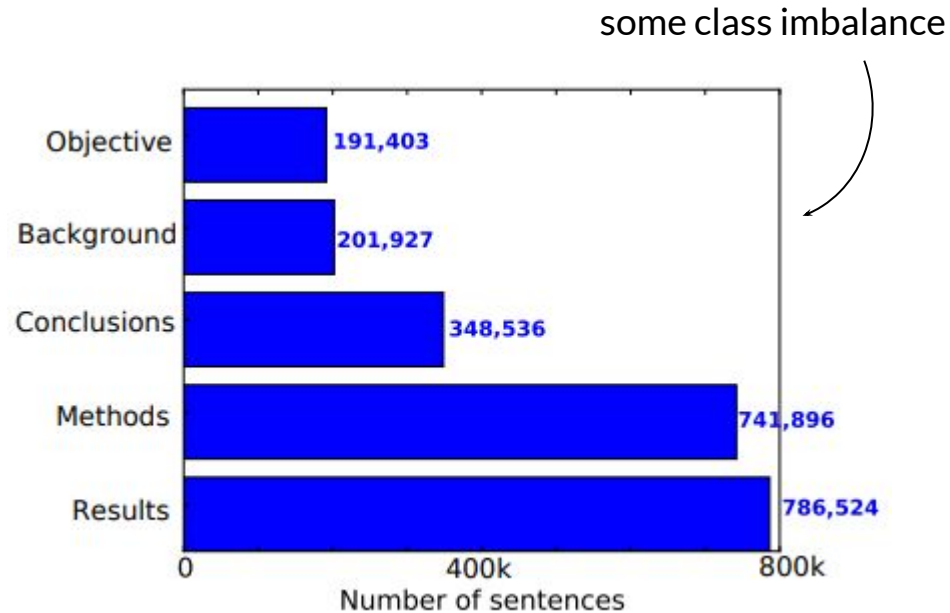


- ~200,000 PubMed abstracts of randomized controlled trials, totaling 2.3 million sentences.
- Designed for sequential sentence classification  Useful to facilitate literature reviewing.
Labels: background, objective, method, result, conclusion.
- Other interesting use cases:
 -  automatic text summarization
 -  information extraction, e.g. what is the scientific claim in this abstract?
 -  information retrieval, e.g. what is the effect of X drug on Y cancer patients?
- 3 data splits: **Train** (model training); **Dev** (hyperparameter selection); **Test** (final model testing).

Data Format & Preprocessing Example



Sequential Sentence Classification



- Multiclass classification
- **Evaluation metric:** F1-score (weighted).
- Analyse the **confusion matrix** between output classes obtained with your best-performing model.

TASK 1: BASELINE MODEL

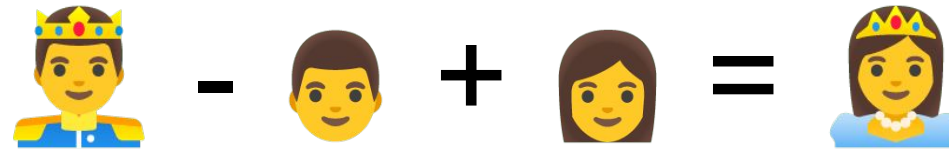
- **Preprocessing:** lowercasing, stop-words & punctuation removal etc.
- Obtain sentence embeddings through **TF-IDF** (e.g. `sklearn TfidfVectorizer`).
- Train classifier to predict corresponding abstract class.

documents: sentences or abstracts
in this task?

Source: Deroncourt & Lee, IJCNLP 2017.

Word Embeddings

How can we translate the *lexical + semantic meaning of words* to a *numerical entity*?



king - man + woman = queen

The diagram shows a visual equation using emojis. On the left, a king emoji (a man with a crown) is followed by a minus sign, a man emoji (a man's face), a plus sign, a woman emoji (a woman's face), and an equals sign. To the right of the equals sign is a queen emoji (a woman with a crown). Below the emojis, the text 'king - man + woman = queen' is written.

Word2Vec: word embeddings depends on **context**.

CBOW: predict word from context. *Skip N-gram:* predict context from word.

FastText: similar but with sub-word decompositions. Allows to embed unseen words.

TASK 2: WORD EMBEDDING

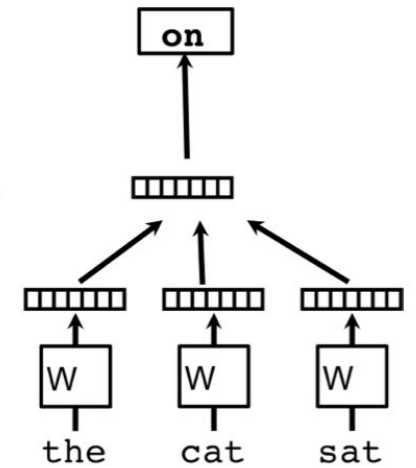
- Train a word embedding model such as **Word2Vec** or **FastText** (we recommend using the `gensim` library).
- Obtain sentence embeddings by averaging or concatenating word embeddings.
- Train classifier to predict corresponding abstract class.

Optional: discuss any interesting semantic relationships between word embeddings.

Classifier

Average/Concatenate

Word Matrix



Transformer-Based Language Models

Recap on **BERT** (Bidirectional Encoder Representations from Transformers, Devlin et al., 2018):

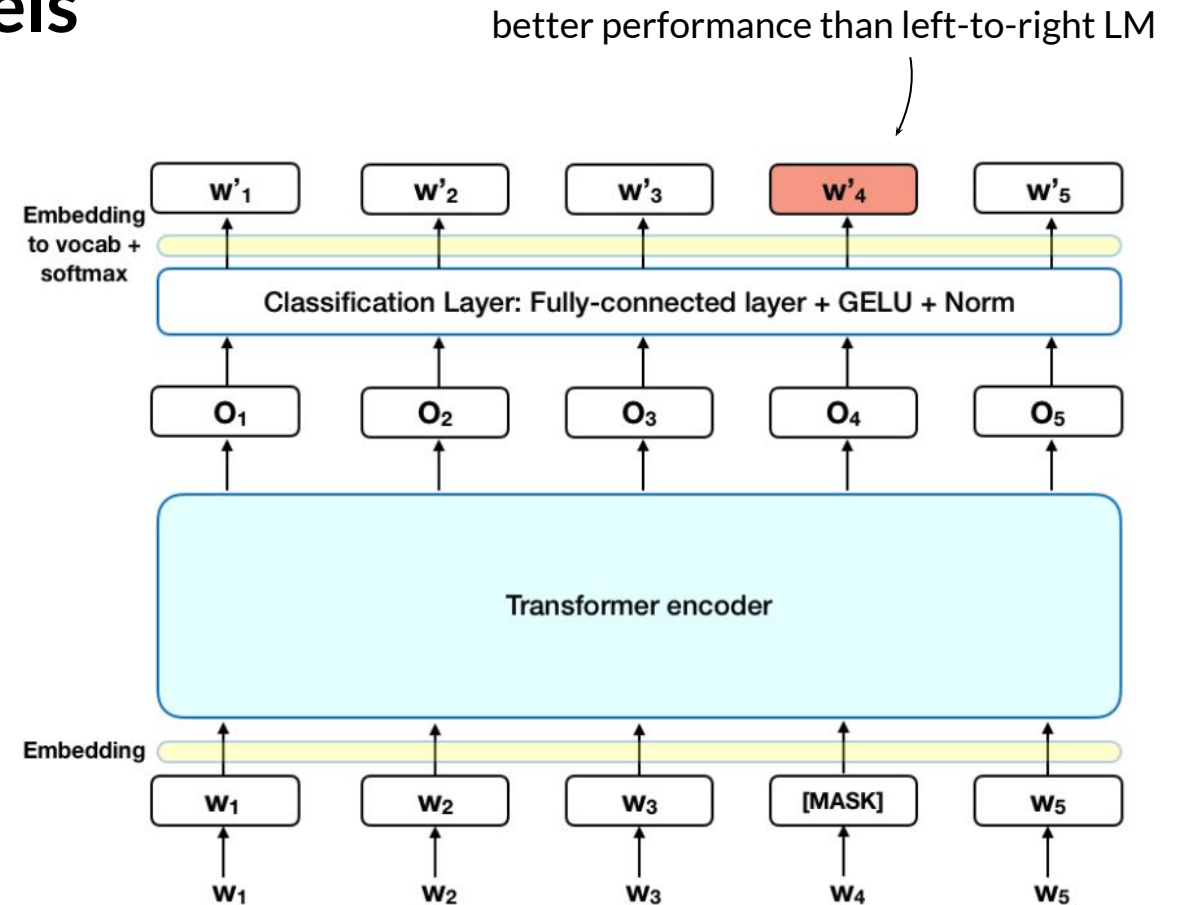
- Encoder of Transformer architecture.
- Trained via Masked LM (right) and Next Sentence Prediction.

BERT pre-trained on biomedical language can be found on HuggingFace. Think about what data the model should be pre-trained on (MIMIC, PubMed, etc.).

TASK 3: TRANSFORMER MODEL

Evaluate the performance of **BERT** on the given task. See mini-tutorial in additional slides.

- Pre-trained, no fine-tuning.
- Pre-trained and fine-tuned on our dataset. Which parameters/layers must be tuned?



Deliverables

- Solve all tasks.
- **Report** of max. 4 pages, 11pt (+ 1 page for references + 1 page of appendix if needed).
- **Well-commented code/jupyter notebooks** with conda environment and README.
- Do not hardcode any results! We will run your code.
- Ensure sequential execution and **reproducibility**.
- **Do not copy solutions from previous projects!** We are aware of all existing solutions on github. We run code similarity checks and check for plagiarism in the reports from previous years solutions. Any plagiarism will result in a 0 grade for all projects.
- **Deadline: 25.04.2022**

Grade

- To grade the project we will focus (on equal parts) on:
 - the content, organisation, clarity, quality and writing of the **final report**.
 - the quality of the **implementation** (reproducibility and clarity).
 - the **creativity/performance*** of the methods used to solve the tasks, and the reasons behind the choices.
- The **prerequisites** to get the maximum grade are:
 - write a clear and **good report**.
 - submit a **clean code** with **easy instructions** on how to reproduce each result of the report.
 - solve every task with **well-justified** methods.
 - **bonus: implement creative models for one task (e.g. alternative embedding or language model)**

*We will consider resource constraints. Aside from correct baseline implementation, the aim is not to get the best performance but to explore and discuss relevant methods.

Questions?

Also on Moodle (preferred: your classmates probably have similar questions!)
or by email at alizee.pace@ai.ethz.ch.

Additional Slides

TF-IDF

In **TF**, all terms are equally important. **TF-IDF** scales down the weight of frequently-occurring terms.
via Document Frequency (**DF**) = (# of documents that contain a term t) / (total # of documents)

- high TF-IDF weight when term occurs *many times* within *few* documents → highly relevant!
- lower TF-IDF weight when term occurs *rarely* or in *many* documents.

Useful to extract essential **keywords** from text documents.

Pre-trained BERT model

Full tutorial on [HuggingFace](#) 🙌
(+ other *How-to* guides)

which AutoModel should you use? i.e. what task are we solving?

```
from transformers import AutoTokenizer, AutoModel
```

```
tokenizer = AutoTokenizer.from_pretrained("emilyalsentzer/Bio_ClinicalBERT")
```

```
model = AutoModel.from_pretrained("emilyalsentzer/Bio_ClinicalBERT")
```

replace with pretrained weights of interest

NB. Pytorch and Tensorflow-compatible interfaces are available if preferred.